一种基于"预读"及简单注意力机制的句子压缩方法 *

鹿忠磊 [1]，刘文芬 [2]，周艳芳 [1]，胡学先 [1]，王彬宇 [1]

(1. 数学工程与先进计算国家重点实验室，郑州 450001; 2. 桂林电子科技大学 计算机与信息安全学院，广西密码学与信息安全重点实验室，广西 桂林 541004)

摘　要：针对英文句子压缩方法进行研究，提出一种基于"预读"及简单注意力机制的压缩方法。在编码器-解码器（encoder-decoder）框架下，以循环门单元（Gated Recurrent Unit，GRU）神经网络模型为基础，在编码阶段对原句语义进行两次建模。首次建模结果作为全局信息，加强二次语义建模，得到更全面准确的语义编码向量。解码阶段充分考虑删除式句子压缩的特殊性，适用简单注意力（3t-Attention）机制，将编码向量中与当前解码时刻最相关的语义部分输入到解码器中，提高预测效率及准确率。在谷歌新闻句子压缩数据集上的实验结果表明，所提压缩方法优于已有公开结果。因此，"预读"及简单注意力机制可有效提高英文句子压缩精度。

关键词：自然语言处理；句子压缩；预读；注意力机制

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.08.0720

Effective method in sentence compression based on Pre-Reading and simple attention mechanism

Lu Zhonglei[1], Liu Wenfen[2], Zhou Yanfang[1], Hu Xuexian[1], Wang Binyu[1]

(1. *State Key Laboratory of Mathematical Engineering & Advanced Computer*, *Zhengzhou* 450001, *China*; 2. *Guangxi Key Laboratory of Cryptogpraphy & Information Security*, *School of Computer Science & Information Security*, *Guilin University of Electronic Technology*, *Guilin Guangxi* 541004, *China*)

Abstract: This paper proposed a method in English sentence compression based on Pre-readin" and Simple Attention Mechanism. On the basis of Gated Recurrent Unit (GRU) and Encoder-Decoder, this paper modeled the original sentence semantics twice in the encoding stage. The first result was used as a global information to strengthen the second semantic model, thus obtaining a more comprehensive and accurate semantic vector. With full consideration of the particularity of the deleted sentence compression, this paper simply adopt the 3t-Attention mechanism in the decoding stage to improve the efficiency and accuracy of prediction, which means that the semantic vectors most relevant to the current decoding time step are inputted to the decoder. The results from the experiments on the Google news sentence compression dataset show that our model significantly outperforms all the recent state-of-the-art methods. Therefore, "Pre-reading" and Simple Attention Mechanism can effectively improve the accuracy of English sentence compression.

Key Words: natural language processing; sentence compression; pre–reading; attention mechanism

0　引言

随着网络信息数量的飞速增长，人们希望精简信息以节约阅读时间。近年来，自然语言处理技术的飞速发展，使计算机逐渐参与至该项工作，句子压缩即是其中重要技术。句子压缩又称句子约简[1]，旨在通过算法处理，模拟人类文本概括和信息提取能力，去除冗余信息，保留核心内容，自动生成合乎语法、语义连贯的精简句，以便读者快速掌握文本要义。该技术广泛用于主题自动提取、摘要自动生成、网络信息搜索、网络舆情监控、推荐系统、问答系统和情感分析等技术中。

传统句子压缩方法通过最小化语法错误比例[2~4]或修剪句法树[5]等得到压缩句子，严重依赖人工设计的规则特征，对专家知识要求较高，且耗费大量的人力物力。而深度学习强大的表示能力，为句子压缩带来新的技术思路。深度学习算法完全

基金项目：国家自然科学基金资助项目（61502527）

作者简介：鹿忠磊（1988-），男，山东泰安市人，硕士研究生，主要研究方向为深度学习与自然语言处理；刘文芬（1965-），女，湖北孝感人，教授，博士，主要研究方向为概率统计理论及应用（liuwenfen@guet.edu.cn）；周艳芳（1993-），女，北京人，硕士研究生，主要研究方向为深度学习与自然语言处理；胡学先（1982-），男，湖北孝感人，讲师，博士，主要研究方向为可证明安全协议和模型；王彬宇（1993），男，山东济南人，硕士研究生，主要研究方向为应用数学.

由数据驱动，自动提取特征，可极大减轻人力物力负担。在各类深度学习范式中，句子压缩属于典型的序列预测任务[6]，即输入原句序列，预测输出压缩句序列。该类任务，通常基于编码器-解码器框架解决，编码器将输入句子序列编码为稠密向量，此向量包含原句语义信息，解码器解码此向量生成原句中各词的保留或删除决策。Fillippova 等人[7]采用类似策略，首次将深度学习模型适用于句子压缩任务，其使用 3 层单向长短时记忆（Long-Short Term Memory，LSTM）网络堆栈作为编码器-解码器组件，在大规模数据集上获得优于传统压缩系统的结果。Tran 等人[8]对 Fillippova 等人的模型结构进行改进，提出一种基于注意力机制（attention mechansim）的双向 LSTM 模型用于句子压缩，且在小数据集上取得较好结果。此外，Sigrid 等人[9]将眼睛跟踪信息（eye-tracking information）纳入句子压缩系统，多任务预测，实现更高的准确性，与 Fillippova 等人工作相似，他们同样使用了三层单向 LSTM 编码器-解码器架构。

但当前相关研究仍存在以下两点不足。一是大部分模型设计过于简单，无法将原句语义充分编码进语义向量（尤其在输入序列较长情况下），致使语义信息丢失严重，无法准确解码。二是常规注意力机制计算复杂度高，注意力信息刻画不直观，在句子压缩任务上适用性不强。

针对以上问题，本文提出类似人类压缩句子行为的“预读”机制，实现更精确句子压缩。首先通读原句，概略掌握全句要义，但此时仅可对句中部分词语作出正确的保留或删除决策；再次逐词阅读，利用通读获得的整体信息调整完善决策，保留重要单词。“预读”机制与该过程类似，两次输入原句序列，并使用首次获得的整体语义信息局部增强第二次语义表示，加大需保留词语语义权重，减小冗余词语语义影响，获取更加全面准确的语义编码向量，从而为解码打下良好基础。同时，在句子压缩任务输入输出序列严格对齐情况下，聚焦与输出序列各时刻预测紧密相关的原句中对应时刻的隐藏层状态，兼顾左右最近邻词对当前词删除或保留决策的较大影响，采用更加科学简单的 3t-Attention 机制，提升解码效率及准确率。本文的主要贡献如下：

a）首次在句子压缩任务上创建使用“预读”机制，获取更加全面准确的语义编码向量；

b）针对句子压缩任务，提出适用更加科学简单的 3t-Attention 机制，降低计算复杂度，提高解码效率及准确率；

c）在多种模型基础上进行了大量对比实验，实验结果对模型及超参数选择具有一定指导意义。

1　预备知识

1.1　GRU 模型

循环神经网络[10]（recurrent neural network，RNN）是常规前馈神经网络(feedforward neural network, FNN)的扩展，该模型允许层内之间的连接和定向循环的出现，能够处理可变长度输入序列[6]，广泛应用于机器翻译[11, 12]、自动问答[13, 14]、语法解析[15, 16]、图像标题生成[17, 18]等。但在实际训练中，常规循环神经网络存在梯度爆炸和梯度消失问题[19]，对长文本表示效果不佳，其两种改进模型——长短时记忆（long short-term memory，LSTM）[20]网络和循环门单元（gated recurrent unit，GRU）[21]网络应运而生。而 Chung 等人[22]研究表明，LSTM 与 GRU 在序列建模任务上表现相当，均通过“门”机制将重要特征保留，保证其在长距离传播的时不被丢弃，但由于 GRU 没有单独存储单元，参数较 LSTM 少，当数据量较大时，在收敛速度和迭代次数上更胜一筹。综合考虑，本文选用 GRU 作为基础模型。GRU 通过更新门(update gate)和重置门(reset gate)的控制，自适应解决 RNN 模型训练中的长程依赖问题。其结构如图 1 所示。

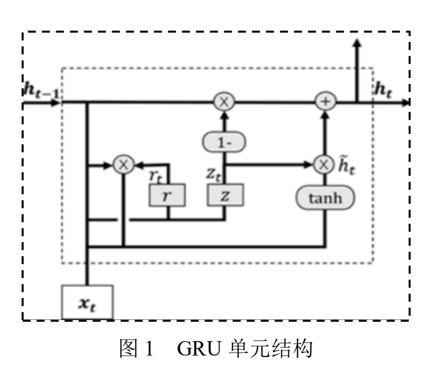


图 1　GRU 单元结构

具体工作原理为

$$z_t = \sigma(W_1 x_t + U_1 h_{t-1} + b_1) \tag{1}$$

$$r_t = \sigma(W_2 x_t + U_2 h_{t-1} + b_2) \tag{2}$$

$$\tilde{h}_t = \tanh(W_3 x_t + U_3 (r_t \odot h_{t-1}) + b_3) \tag{3}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{4}$$

其中：z 表示更新门，r 表示重置门，x 为输入层，h 为隐藏层，$\tilde{h}$ 为中间状态，与 h 对应。重置门 r 决定是否舍弃之前状态，即当 r 趋于 0 时，前一时刻的隐藏层状态信息 h_{t-1} 被忽略，中间状态 $\tilde{h}$ 被重置为当前输入信息。更新门 z 决定是否要将当前时刻隐藏层状态更新为新的中间状态 $\tilde{h}$，即当 z 趋于 1 时，前一时刻的隐藏层状态信息 h_{t-1} 被忽略，当前时刻隐藏层状态被置为中间状态 $\tilde{h}$。更新门和重置门共同决定当前隐藏层输出。U、W 和 b 均为模型参数矩阵，$\odot$ 表示对应元素相乘。

1.2　双向 GRU 模型

无论是 RNN、LSTM，还是 GRU，均只编码利用单向语义信息。进一步，对于时刻 t，其隐藏层输出仅包含 t 时刻之前的信息，即上文信息。而下文信息对整个语义的刻画同样重要。为更好表示整体上下文信息，基于已被成功应用的双向 RNN（Bidirectional RNN，BiRNN）模型[23][24]，提出使用双向 GRU（Bidirectional GRU，BiGRU）模型。该模型可利用历史和将来的所有可用输入信息进行训练，获得更加全面准确的语义向量表示。

与 GRU 相比，BiGRU 使用两个单独的隐藏层双向读取输入：正向和反向。正向以原始顺序（1 到 T）读取输入，反向按照相反顺序（T 到 1）读取输入。时刻 t 的两个隐藏层状态

为

$$\vec{h}_t = \mathrm{GRU}(x_t, \vec{h}_{t-1}) \quad (5)$$

$$\overleftarrow{h}_t = \mathrm{GRU}(x_t, \overleftarrow{h}_{t-1}) \quad (6)$$

BiGRU 的初始状态置为全零向量，即 $\vec{h}_0=0$，$\overleftarrow{h}_{T+1}=0$。根据式(1)~(4)，可计算得到正向隐藏层状态$(\vec{h}_1,\vec{h}_2,\ldots,\vec{h}_T)$和反向隐藏层状态$(\overleftarrow{h}_1,\overleftarrow{h}_2,\ldots,\overleftarrow{h}_T)$，后通过级联方式综合表示语义：

$$h_t = [\vec{h}_t, \overleftarrow{h}_t] \quad (7)$$

从而，隐藏层状态 h_t 同时含上下文信息，可有效提高模型在较长序列上的记忆表现。

1.3　基于双向 GRU 模型的编码器-解码器框架

编码器-解码器框架是自然语言处理问题解决方案的新范式，且被广泛用来解决序列到序列（Sequence to sequence，seq2seq）预测问题[6]，如机器翻译[25][26][27]、自动文摘[28][29][30]等。基于双向 GRU 模型的编码器-解码器框架受 RNN 编码器-解码器的启发，第一部分使用双向 GRU 模型对输入句子序列进行编码，生成一个固定长度的稠密编码向量，该向量包含输入句子序列语义信息。第二部分使用 GRU 模型对编码向量进行解码，逐个预测句中单词的标签。因此，上下文表示（语义编码向量）是编码器-解码器框架下句子压缩任务得以有效解决的关键。

1.4　基于注意力机制的 GRU 模型

Bahdanau 等人[31]提出，在解码产生每个单词时，可使用注意力机制，动态利用输入序列中与各时刻解码相关的具体部分，实现源词与目标词语义对齐，可有效提升模型预测精度。此后，注意力机制广泛应用于学习各种模式之间的对齐，如语音识别任务中语音帧和文本对齐[32]，图像标题生成任务中图像与文本描述对齐[18]等。

基于注意力机制的 GRU 解码器工作原理如下：

$$z_t = \sigma(W_1x_t + U_1h_{t-1} + V_1c_t + b_1 \quad (8)$$

$$r_t = \sigma(W_2x_t + U_2h_{t-1} + V_2c_t + b_2 \quad (9)$$

$$\tilde{h}_t = \tanh(W_3x_t + U_3(r_t \odot h_{t-1}) + V_3c_t + b_3) \quad (10)$$

$$h_t = (1-z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (11)$$

其中：c_t 是基于注意力机制的上下文表示，根据源词与目标词对齐结果动态生成，V 是上下文信息的权重矩阵。Bahdanau 等人[30]将所有编码阶段隐藏层状态的加权和作为 t 时刻的上下文表示，即

$$c_t = \sum_{j=1}^{T} \alpha_{tj}\, h_E^j \quad (12)$$

权重α_{tj}计算如下：

$$r_{tj} = v_\alpha^T \tanh(W_\alpha h_{t-1} + U_\alpha h_E^j) \quad (13)$$

$$\alpha_{tj} = softmax(r_{tj}) \quad (14)$$

本文称该种表示方法为常规注意力机制。

2　本文模型

本文提出的方法属于删除式句子压缩范畴，即保留重要单词，删除冗余单词。该过程可表示为为原句中每个单词标注 0/1 标签问题，0 代表删除，1 代表保留。进一步，该任务由“输入输出均为单词序列”转换为“输入为单词序列、输出为 0/1 序列”问题。如下例所示：

> 输入句：A woman from Lycoming County has been charged with theft from her place of work.
> 压缩句：A woman has been charged with theft.
> 输　出：1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1

本文基于 Sutskever 等人[6]提出的序列到序列范式处理该问题，基本思想即采用端到端策略训练模型，使输入句子对应正确输出的概率最大。具体而言，对于每个训练样本 (X,Y)，通过随机梯度下降法（Stochastic Gradient Descent，SGD）求解以下优化问题，学习模型参数 θ^*为

$$\theta^* = arg\max_{\theta} \sum_{X,Y} logp(Y|X;\theta) \quad (15)$$

总和为所有训练样本预测损失的加和。使用链式法则对概率 p 建模，得到

$$p(Y|X;\theta) = \prod_{t=1}^{T} p(Y_t|Y_1,\ldots,Y_{t-1},X;\theta) \quad (16)$$

此处无任何独立性假设。得到最优参数 θ^* 后，即可估计压缩结果为

$$\hat{Y} = arg\max_{Y} logp(Y|X;\theta^*) \quad (17)$$

2.1　“预读”机制

对人类来说，如不“预读”输入句，即在不掌握全句信息情况下，将很难得到一个词语或一句话的正确表示，继而影响句子压缩准确率。同样，各类循环神经网络虽在序列建模方面表现不俗，但 t 时刻的隐藏层状态仅依赖于历史信息，且双向模型下双向隐藏层状态之间缺乏直接互动，势必导致压缩效果不佳。

本文“预读”机制背后思想十分直观，一般人类压缩句子时，首先通读原句，即文中所指“预读”，经“预读”后，获得整句语义，在此基础上，再次读取原句，对各词逐一作出删除或保留决策。对计算机而言，该过程的实现流程为，将原句输入神经网络，学习得到句子的稠密分布式表示，即语义向量，该语义向量用于衡量原句中词语重要性，获得原句中各词语义权重；将原句再次输入神经网络，使用首次语义建模权重对语义特征进行再提取，突出高信息量词的语义贡献，削弱非保留词影响，实现语义表达更具侧重，服务任务目标。

其概略流程如图 2 所示。

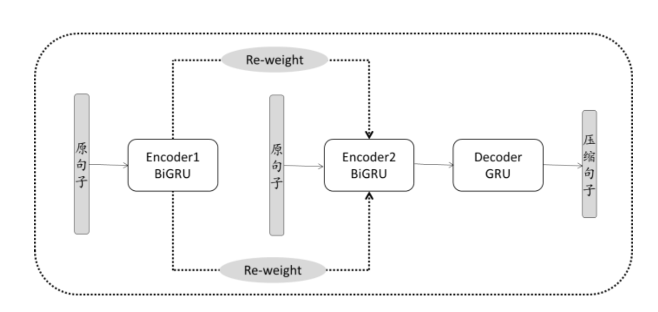


图 2　“预读”机制流程图

以正向“预读”为例（反向“预读”可同理推广），假设输入序

列为$(x_1,x_2,\dots,x_T)$，使用标准 GRU 首次读取输入序列，

$$\vec{h}_t^1 = GRU^1(x_t,\vec{h}_{t-1}^1) \tag{18}$$

其中，GRU^l 定义如式(1)～(4)。获得整个句子特征向量 h_T^1 后，计算权重向量 α_t，该向量辅助语义二次建模。其具体工作原理如图 2(b)所示。若当前单词 x_t 对应权重 α_t 趋近于 0，则二次编码时对应隐藏层状态 h_t^2 所携带的信息几乎全部来自于上一时刻的隐藏层状态 h_{t-1}^2，而忽略当前输入词 x_t 的影响；若权重 α_t 接近于 1，则该结构与标准 GRU 相似，仅受当前词的影响。二次建模规则为

$$\vec{h}_t^2 = (1-\alpha_t)\odot\vec{h}_{t-1}^2 + \alpha_t\odot GRU^2(x_t,\vec{h}_{t-1}^2) \tag{19}$$

其中 α_t 计算如下：

$$\alpha_t = \sigma(W\vec{h}_t^1 + U\vec{h}_T^1 + Vx_t) \tag{20}$$

矩阵 W，U 和 V 是模型参数，σ 是 Sigmoid 函数。α_t 是一个权重向量，与隐藏层状态维度相同，代表隐藏层状态每一维的重要程度。此处使用向量而非数值，究其原因是隐藏层状态不同维度代表不同的语义语法特征，单值权重无法捕捉各维信息及重要程度变化。

将公式中 $GRU^2(x_t,\vec{h}_{t-1}^2)$ 进一步展开：

$$GRU^2(x_t,\vec{h}_{t-1}^2) = (1-z_t^2)\odot\vec{h}_{t-1}^2 + z_t^2\odot\tilde{\vec{h}}_t^2 \tag{21}$$

将上式代入(19)得

$$\vec{h}_t^2 = (1-\alpha_t)\odot\vec{h}_{t-1}^2 + \alpha_t\odot((1-z_t^2)\odot\vec{h}_{t-1}^2 + z_t^2\odot\tilde{\vec{h}}_t^2) \tag{22}$$

简化：

$$\vec{h}_t^2 = \left(1-\alpha_t\odot z_t^2\right)\odot\vec{h}_{t-1}^2 + (\alpha_t\odot z_t^2)\tilde{\vec{h}}_t^2 \tag{23}$$

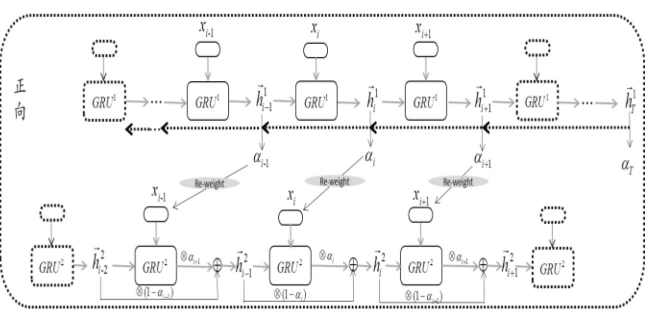


图 3 “预读”机制原理图

2.2 nt-Attention 模型

2.2.1 基于 t-Attention 的 GRU 模型

常规注意力机制计算复杂度高，注意力信息重复冗余。在删除式句子压缩任务中，由于是将句子序列转换为 0/1 序列，因此，输入序列与输出序列等长且严格对齐，该模式下采用常规注意力机制的必要性不大。Tran 等人[8]提出，将编码阶段各时刻隐藏层状态 h_E^t 直接作为解码器端对应时刻的注意力信息，即只考虑与被预测词最相关的上下文信息，而非关注句子所有组成部分，从而有效去除冗余信息。

$$h_t = f(x_t,y_{t-1},h_E^t) \tag{24}$$

但本文认为，最近邻词对当前词的删除及保留决策影响较大，因此，考虑将 t-Attention 扩展，通过 2t-Attention、3t-Attention，增强注意力语义信息，降低决策错误率。

2.2.2 基于 2t-Attention 的 GRU 模型

以正向 2t-Attention 为例，预测单词 x_t 的标签时，使用编码阶段 x_{t-1} 和 x_t 对应的隐藏层状态组合作为上下文语义表示输入解码器，

$$\vec{h}_t = f(x_t,y_{t-1},[\vec{h}_E^{t-1},\vec{h}_E^t]) \tag{25}$$

其中：$[h_E^{t-1},h_E^t]$ 表示编码阶段 t-1 和 t 时刻隐藏层状态的级联。

类似，反向 2t-Attention 可表示为

$$\overleftarrow{h}_t = f(x_t,y_{t-1},[\overleftarrow{h}_E^t,\overleftarrow{h}_E^{t+1}]) \tag{26}$$

2.2.3 基于 3t-Attention 的双向 GRU 模型

为综合考虑最近邻词对当前词的影响，将 2t-Attention 进一步扩展为 3t-Attention，即

$$h_t = f(x_t,y_{t-1},[\vec{h}_E^{t-1},\vec{h}_E^t,\overleftarrow{h}_E^t,\overleftarrow{h}_E^{t+1}]) \tag{27}$$

具体架构如图 3 所示。输出层为 Softmax 分类器，预测对应单词或符号的标签，输出一个 3 维独热（one-hot）向量：若保留，向量第一维为 1，代表标签 1；若删除，向量第二维为 1，代表标签 0；若为句末结束字符，向量第三维为 1，指示解码预测开始。

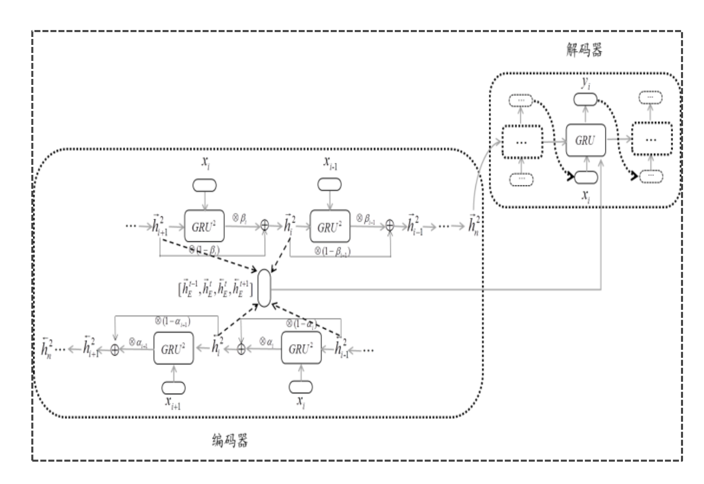


图 3 基于 3t-Attention 的双向 GRU 模型

3 数据与实验

3.1 数据与预处理

深度学习模型包含大量参数，其训练需要充足数据。针对句子压缩平行数据匮乏问题，Filippova 等人[5]提出一种自动生成句子压缩数据集的新方法，构造了大量来自谷歌新闻（Google newswire）的“原句-压缩句”对。本文基于其公开的 4 万对数据，进行实验与对比实验。数据集划分为 3 部分，其中 36000 对作为训练集，2000 对作为验证集，剩余 2000 对作为测试集。

实验前，对数据进行预处理。使用 NLTK[1]分词工具对原句

---

[1] http://www.nltk.org/

进行分词，然后借助 word2vec[33]模型进行预训练，获得 97 维词向量。经验证，预训练不仅能加速训练过程，且训练所得模型效果优于随机初始化，因此本文实验均在预训练基础上实施。选取前 8000 个高频词组成训练词表，超出词表的单词或字符统一由“unk”（unknown）表示。在每个句子结尾，添加一个特殊符号“eos”（end of sentence），作为解码启动指示。基于以上处理，构造标准标签序列，对句子中每个单词进行单独标记，若单词被保留，标注为 1；若单词被删除，标注为 0；若单词是“eos”，标注为 2。

3.2 实验设置

实验中，编码器和解码器端的隐藏层单元数设置为 100。输入为 100 维向量，其中前 97 维为当前输入单词的向量表示。后 3 维在编码和解码阶段不同，编码阶段为全零向量，解码阶段为前一个单词标准标签（训练期间）或预测标签（测试期间）的独热向量表示。

根据 Greff 等人[34]关于参数设定研究经验，结合本文数据量，将学习率初始化为 0.001，每 1000 个训练步的衰减率[35]设定为 0.9。采用提前结束（early-stop）策略[36]，即当验证集的 F1 分数不增加达 5 轮后，系统结束训练，进一步防止过拟合。具体模型参数设置如表 1 所示。

表 1　模型参数设置

参数	设定值	实验范围
最大句子长	120	120
词表大小	8000	16000
词向量维度	97	100,150,200
隐藏层大小	100	100,150,200
最大轮数	50	50
批大小	1000	1000,2000
退出率	0.7	0.7
学习率	0.001	0.001
衰减率	0.9	0.9

选用 Theano[2] 作为本文实验基础框架，其他环境配置 Intel core i7 处理器，16GB 内存，64 位 Ubuntu 16.04 LTS 操作系统。由于模型训练过程中计算量较大，为提高训练效率，在此基础上额外增加一块 GTX 1070 GPU 加速卡。

3.3 实验结果

与 Tran 等人[8]工作一致，本文使用 F1 分数和 per-sentence accuracy 值（简记为 Acc）两个指标进行评估，Acc 指完全再现的压缩句子所占比例。基于以上设置，从基准模型、注意力范围和有无“预读”机制等多个角度进行大量对比实验，并考察了部分超参数影响，系统给出了有关结论。

为便于表述，对模型构成作以下规定，“F”表示正向，“B”表示反向，“Bi”表示双向，“R”表示“预读”，“tA”表示 t-Attention，则“BiR-3tA”即表示“双向预读 3t-Attention”模型。

表 2　基准模型

模型	F1	Acc
3LSTM[7]	0.7445	0.225
BiLSTM-tA[8]	0.7681	**0.315**
BiGRU-tA	**0.7682**	0.314

表 2 对比了各种基准模型，第一行和第二行分别为 Fillippova 等人[7]和 Tran 等人[8]提出的模型。结果表明，在现有数据集上，BiLSTM-tA 与 BiGRU-tA 模型均可取得较 3LSTM 模型更好结果，且效果相当，凸显出双向模型及注意力机制对文本语义建模的提升作用。下面以 BiGRU-tA 为基础模型，围绕注意力范围与有无“预读”机制两个方面，逐步增加模型复杂度，对各种组合模型进行研究分析。

表 3　注意力范围

模型	F1	Acc
BiGRU-tA	0.7683	0.314
BiGRU-F2tA	0.7721	0.316
BiGRU-B2tA	0.7745	0.317
BiGRU-3tA	**0.7786**	**0.320**

表 3 在无“预读”机制情况下，对比注意力范围对实验结果的影响。其中，第二行表示将当前词及其前一个词的联合上下文表示作为注意力信息，第三行表示将当前词及其后一个词的联合上下文表示作为注意力信息，且后者效果优于前者，该结论与文献[5]结论一致，即倒序输入较正序输入解码准确率会有所提高。该表结果显示，附加最近邻词语义表示模型效果均优于 BiGRU-tA 模型，且同时考虑前后最近邻词语义表示的 3t-Attention 模型取得最好效果。可以看出，左右最近邻词对当前词的删除决策影响较大。在扩大注意力范围的同时，本文考察了级联及加和两种刻画上下文语义信息的方式，结果表明，级联效果优于加和。

表 4　有无及单双向“预读”机制

模型	F1	Acc
BiGRU-3tA	0.7786	0.320
FR-3tA	0.7822	0.322
BR-3tA	0.7861	0.327
BiR-3tA	**0.7936**	**0.329**

表 4 在 3t-Attention 模型基础上，对比有无“预读”及单双向“预读”机制对实验结果的影响。从表中可以看出，带有“预读”机制的模型效果好于无“预读”机制模型，且后向“预读”模型好于前向“预读”模型，双向“预读”模型取得最好结果，F1 值高达 0.7936。

[2] http://deeplearning.net/software/theano/

此外，进一步对部分超参数进行调优，如词向量维度、隐藏单元数和词表大小等。如表 5 所示，在当前数据量下，适当增大词向量维度和隐藏层单元数，模型效果会有所提升，而词表大小对实验结果影响较小。

表 5　部分超参数影响

模型	词向量维度	隐藏单元数	词表大小	F1	Acc
BiR-3tA	100	100	8000	0.7936	0.329
	150	100	8000	0.7954	0.331
	200	100	8000	0.7966	0.330
	100	150	8000	0.7945	0.329
	100	200	8000	0.7982	0.332
	100	100	16000	0.7935	0.329

综上，在句子压缩任务中，GRU 模型可代替 LSTM 模型，“预读”机制能较好提升模型文本建模能力，同时，融合左右最近邻词语义表示的注意力机制简单有效。“预读”机制在双向 GRU 模型基础上，通过两次语义建模，模拟人类阅读行为，利用全局信息进行局部调整，增大高信息量词在语义表示中的权重，可提高压缩精度。简单注意力机制利用当前词的直接上下文表示，并附加左右最近邻词语义，忽略冗余信息，避免模型受到语法错误影响，可提高压缩效率和精度。此外，模型中超参数的选择影响其性能表现，但具有一定的经验性。

3.4　示例分析

表 6 显示了不同模型下句子压缩情况。可以看出，本文提出的 BiR-3tA 模型在大多数情况下，压缩内容完整丰富，语法正确合理。使用短输入句进行测试时（前两个），三个模型均可得到正确压缩；压缩长句时，本文模型优于其他压缩系统。

其中 3LSTM 模型表现较差，究其原因，该模型包含约 100 万个参数，使用当前训练数据集（36000 个句子对），不足以将参数调整至最优。而本文提出的模型，参数较少，在强化文本语义建模基础上，能够高效捕捉数据分布规律，提取频繁出现特征，即关键语义、语法等信息，融合简单注意力机制，聚焦与当前词紧密相关的上下文影响信息，采用端到端的联合训练，自适应学习句子压缩决策因子，为输入文本中的每个单词分配一个有意义的权重，突出重要的动词和名词，而忽略常用的单词，如介词等，实现更精准压缩。

4　结束语

本文针对英文句子压缩任务，提出一种基于“预读”及简单注意力机制的压缩方法，在编码器-解码器框架下对原句语义进行两次建模，首次建模结果作为全局信息，调整第二次建模结果，获取更加全面准确的语义特征，并利用简单 3t-Attention 机制，聚焦编码阶段与当前解码时刻最相关的上下文信息，附加左右最近邻词语义影响，忽略冗余信息，在简化计算基础上，提高解码预测准确率.在未使用任何人工设计特征的情况下，本文提出的模型在谷歌新闻压缩数据集上表现突出，F1 值高达 0.7936。下一步，采用生成式模型，研究直接生成压缩句子的方法。

表 6　不同模型下句子压缩情况

输入句子： Gubernatorial candidate Abbott calls for greater privacy protections, legalizing open carry.

标准压缩： Abbott calls for greater privacy protections.

3LSTM： candidate Abbott calls for privacy protections.

BiLSTM-tA： candidate Abbott calls for greater privacy protections.

BiR-3tA： Abbott calls for greater privacy protections.

输入句子： Fun and food are two ways to help Children's Hospital of Illinois during the month of July.

标准压缩： Fun and food are two ways to help Children's Hospital.

3LSTM： Fun and food are ways to help Children's Hospital.

BiLSTM-tA： Fun and food are two ways to help Children's Hospital.

BiR-3tA： Fun and food are two ways to help Children's Hospital.

输入句子： Eurozone business activity slowed in October, coming off a 27-month high in September to highlight concerns the economy is recovering only slowly from recession, a survey showed on Thursday.

标准压缩: Eurozone business activity slowed.

3LSTM: Eurozone business activity slowed, coming off a high in September economy is recovering slowly.

BiLSTM-tA: Eurozone business activity slowed, coming off a 27-month high.

BiR-3tA： Eurozone business activity slowed.

输入句子： Eka Software Solutions, the fast growing global provider of end-to-end commodity management software, today announced that GrainCorp has gone live in Australia on Eka's commodity management platform, as the rst stage of a global implementation.

标准压缩: GrainCorp has gone live on Eka's commodity management platform.

3LSTM: GrainCorp has gone live in Australia.

BiLSTM-tA: GrainCorp has gone live on platform.

BiR-3tA： GrainCorp has gone live on Eka's commodity management platform.

参考文献：

[1] 陈劲光，何婷婷，李芳等. 基于概率和句法分析的中文句子修剪 [C]// 第五届全国青年计算语言学研讨会论文集, 2010.

[2] Jing H. Sentence reduction for automatic text summarization [C]// Proc of Applied Natural Language Processing Conference. 2000: 310-315.

[3] 景秀丽，郑学伟. 基于 Noisy-Channel Model 的句子压缩方法 [J]. 电大理工, 2005 (2): 39-41.

[4] Clarke J, Lapata M. Global inference for sentence compression: An integer linear programming approach [J]. Journal of Artificial Intelligence Research, 2008, 31: 399-429.

[5] Filippova K, Altun Y. Overcoming the lack of parallel data in sentence compression [C]// Proc of EMNLP. 2013: 1481-1491.

[6] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]// Advances in Neural Information Processing Systems. 2014: 3104-3112.

[7] Filippova K, Alfonseca E, Colmenares C A, et al. Sentence compression by deletion with LSTMs [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2015: 360-368.

[8] Tran N T, Luong V T, Nguyen N L T, et al. Effective attention-based neural architectures for sentence compression with bidirectional long short-term memory [C]// Proc of the 7th Symposium on Information and Communication Technology. New York: ACM Press, 2016: 123-130.

[9] Klerke S, Goldberg Y, Søgaard A. Improving sentence compression by learning to predict gaze [J]. arXiv preprint arXiv: 1604. 03357, 2016.

[10] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model [C]// Proc of Interspeech. 2010, 2: 3.

[11] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [J]. arXiv preprint arXiv: 1508. 04025, 2015.

[12] Ling W, Trancoso I, Dyer C, et al. Character-based neural machine translation [J]. arXiv preprint arXiv: 1511. 04586, 2015.

[13] Wang S, Jiang J. Machine comprehension using match-lstm and answer pointer [J]. arXiv preprint arXiv: 1608. 07905, 2016.

[14] Trischler A, Ye Z, Yuan X, et al. A parallel-hierarchical model for machine comprehension on sparse data [J]. arXiv preprint arXiv: 1603. 08884, 2016.

[15] Legrand J, Collobert R. Joint RNN-based greedy parsing and word composition [J]. arXiv preprint arXiv: 1412. 7028, 2014.

[16] Vinyals O, Kaiser Ł, Koo T, et al. Grammar as a foreign language [C]// Advances in Neural Information Processing Systems. 2015: 2773-2781.

[17] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.

[18] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention [C]// Proc of International Conference on Machine Learning. 2015: 2048-2057.

[19] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult [J]. IEEE Trans on Neural Networks, 1994, 5 (2): 157-166.

[20] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[21] Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches [J]. arXiv preprint arXiv: 1409. 1259, 2014.

[22] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv: 1412. 3555, 2014.

[23] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Trans on Signal Processing, 1997, 45 (11): 2673-2681.

[24] Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM [C]// Proc of IEEE Workshop on Automatic Speech Recognition and Understanding. 2013: 273-278.

[25] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. arXiv preprint arXiv: 1406. 1078, 2014.

[26] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [J]. arXiv preprint arXiv: 1508. 04025, 2015.

[27] Cohn T, Hoang C D V, Vymolova E, et al. Incorporating structural alignment biases into an attentional neural translation model [J]. arXiv preprint arXiv: 1601. 01085, 2016.

[28] Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization [J]. arXiv preprint arXiv: 1509. 00685, 2015.

[29] Hu B, Chen Q, Zhu F. Lcsts: A large scale chinese short text summarization dataset [J]. arXiv preprint arXiv: 1506. 05865, 2015.

[30] Nallapati R, Zhou B, Gulcehre C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond [J]. arXiv preprint arXiv: 1602. 06023, 2016.

[31] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv: 1409. 0473, 2014.

[32] Chorowski J, Bahdanau D, Cho K, et al. End-to-end continuous speech recognition using attention-based recurrent NN: first results [J]. arXiv preprint arXiv: 1412. 1602, 2014.

[33] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. 2013: 3111-3119.

[34] Greff K, Srivastava R K, Koutník J, et al. LSTM: a search space odyssey [J]. IEEE Trans on Neural Networks and Learning Systems, 2016.

[35] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15 (1): 1929-1958.

[36] Raskutti G, Wainwright M J, Yu B. Early stopping and non-parametric regression: an optimal data-dependent stopping rule [J]. Journal of Machine Learning Research, 2014, 15 (1): 335-366.